R Notebook

```
data1<-read.table("/Users/phamkhoa/Documents/university/3/stats_model_1/w2/canoladiesel.txt", sep="\t",
attach(data1)
```

```
# Problem 1

#a
model.inv1.ex1<-glm(Yield~Time, family = gaussian(link = "inverse"), data = data1)
summary(model.inv1.ex1)
```

```
##
## Call:
## glm(formula = Yield ~ Time, family = gaussian(link = "inverse"),
##     data = data1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -46.845   -7.784   -3.919    3.123   50.755
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.145935   0.093478   1.561    0.137
## Time        -0.002809   0.002085  -1.347    0.196
##
## (Dispersion parameter for gaussian family taken to be 667.9073)
##
##     Null deviance: 17848  on 18  degrees of freedom
## Residual deviance: 11355  on 17  degrees of freedom
## AIC: 181.39
##
## Number of Fisher Scoring iterations: 7
```

```
newdata <-data.frame(Time = 40)
predict(model.inv1.ex1, newdata = newdata, type = "response")
```

```
##        1
## 29.79662
```

```
#b
model.inv1gamma.ex1<-glm(Yield~Time, family = Gamma(link = "inverse"), data = data1)
model.identity1gamma.ex1<-glm(Yield~Time, family = Gamma(link = "identity"), data = data1)
model.log1gamma.ex1<-glm(Yield~Time, family = Gamma(link = "log"), data = data1)


AIC(model.identity1gamma.ex1)
```

```
## [1] 152.502
```

```
AIC(model.inv1gamma.ex1)
```

```
## [1] 153.5631
```

```
AIC(model.log1gamma.ex1)
```

```
## [1] 151.7177
```

```
# choose (i) log link model
```

```
#c
model.log123gamma.ex1<-glm(Yield ~ Time + Temp + Methanol, family = Gamma(link = "log"), data=data1)
summary(model.log123gamma.ex1)
```

```
##
## Call:
## glm(formula = Yield ~ Time + Temp + Methanol, family = Gamma(link = "log"),
##     data = data1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7117  -0.3192  -0.0254   0.2524   0.4851
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.402855   2.009810  -8.659 3.21e-07 ***
## Time          0.066847   0.007972   8.385 4.80e-07 ***
## Temp          0.065005   0.007776   8.360 4.99e-07 ***
## Methanol      1.008914   0.239162   4.219 0.000745 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.170404)
##
##     Null deviance: 28.3065  on 18  degrees of freedom
## Residual deviance:  2.7836  on 15  degrees of freedom
## AIC: 121.56
##
## Number of Fisher Scoring iterations: 8
```

```
newdata <-data.frame(Time = 40, Temp = 260, Methanol = 1.75)
pred<-predict(model.log123gamma.ex1, type="response", level=0.95, newdata=newdata, se.fit = TRUE)

eta<-predict(model.log123gamma.ex1, type="link", newdata=newdata, se.fit = TRUE)
link.lowerbound<-eta$fit-qnorm(0.975)*eta$se.fit
link.upperbound<-eta$fit+qnorm(0.975)*eta$se.fit

upper<-exp(link.upperbound) ### !!!!
lower<-exp(link.lowerbound) ### !!!!
upper
```

```
##        1
## 68.33616
```

```
lower
```

```
##        1
## 38.51991
```

```
# 95% CI = (38.51991, 68.33616)
```

```
#d
gammaH0 <- glm(Yield ~ Time + Temp, family = Gamma(link = "log"), data=data1)
test<-anova(gammaH0, model.log123gamma.ex1, test = "F")
test$`Pr(>F)`
```

```
## [1]          NA 0.001575748
```

```
#e
model.e.N <- glm(Yield ~ Time + Temp + Methanol, family = gaussian(link = "log"), data=data1)
model.e.G <- glm(Yield ~ Time + Temp + Methanol, family = Gamma(link = "log"), data=data1)
model.e.IG <- glm(Yield ~ Time + Temp + Methanol, family = inverse.gaussian(link = "log"), data=data1)

AIC(model.e.N)
```

```
## [1] 142.6993
```

```
AIC(model.e.G)
```

```
## [1] 121.5555
```

```
AIC(model.e.IG)
```

```
## [1] 128.6878
```

```
## MSE

mean(residuals(model.e.N, type="pearson")^2)
```

```
## [1] 63.19944
```

```
mean(residuals(model.e.G, type="pearson")^2)
```

```
## [1] 0.1345295
```

```
mean(residuals(model.e.IG, type="pearson")^2)
```

```
## [1] 0.018071
```

```
shapiro.test(residuals(model.e.N, type="pearson"))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model.e.N, type = "pearson")
## W = 0.83734, p-value = 0.004203
```

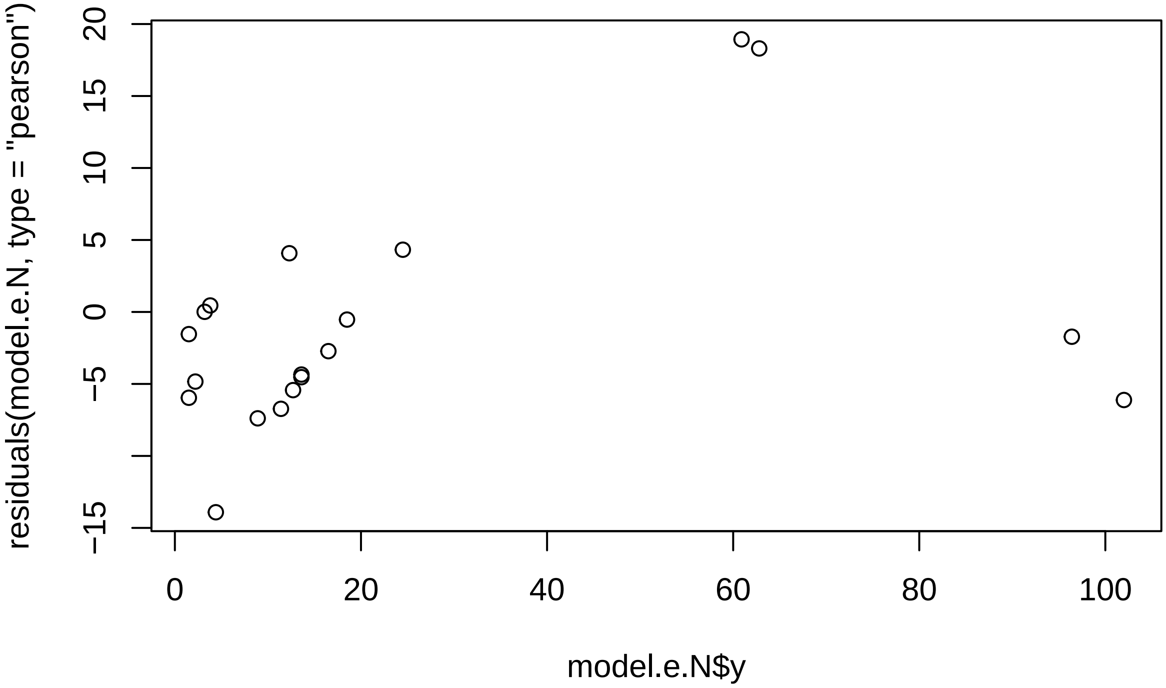
```
shapiro.test(residuals(model.e.G, type="pearson"))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model.e.G, type = "pearson")
## W = 0.93471, p-value = 0.2114
```

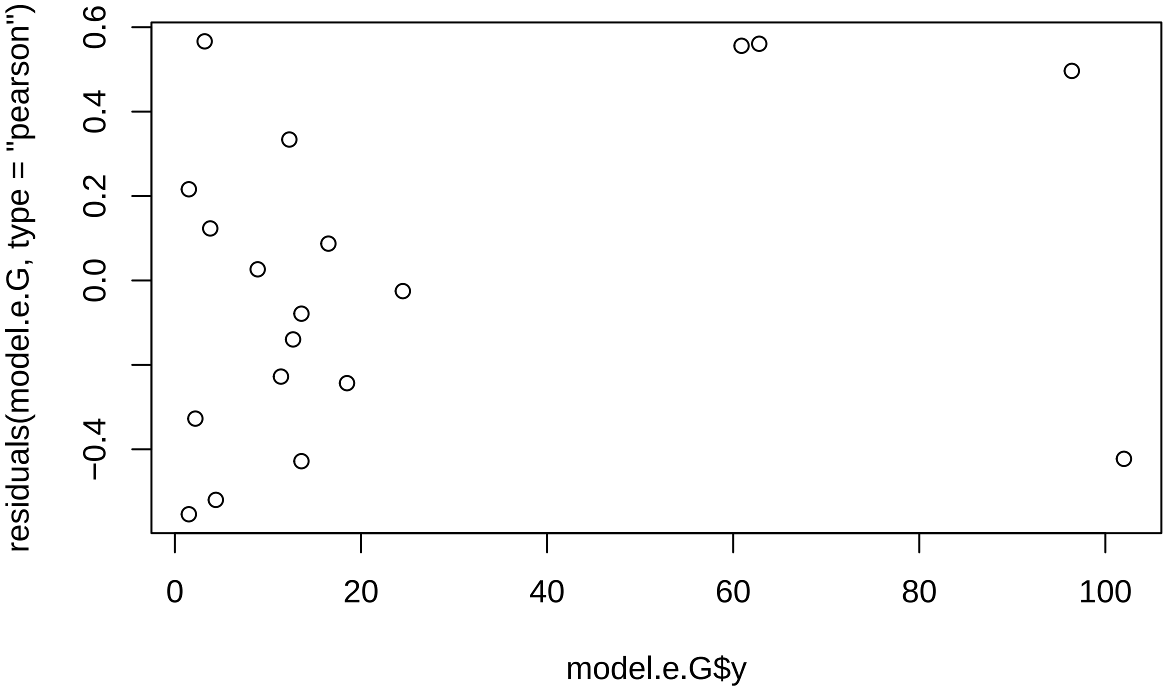
```
shapiro.test(residuals(model.e.IG, type="pearson"))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model.e.IG, type = "pearson")
## W = 0.98053, p-value = 0.9483
```

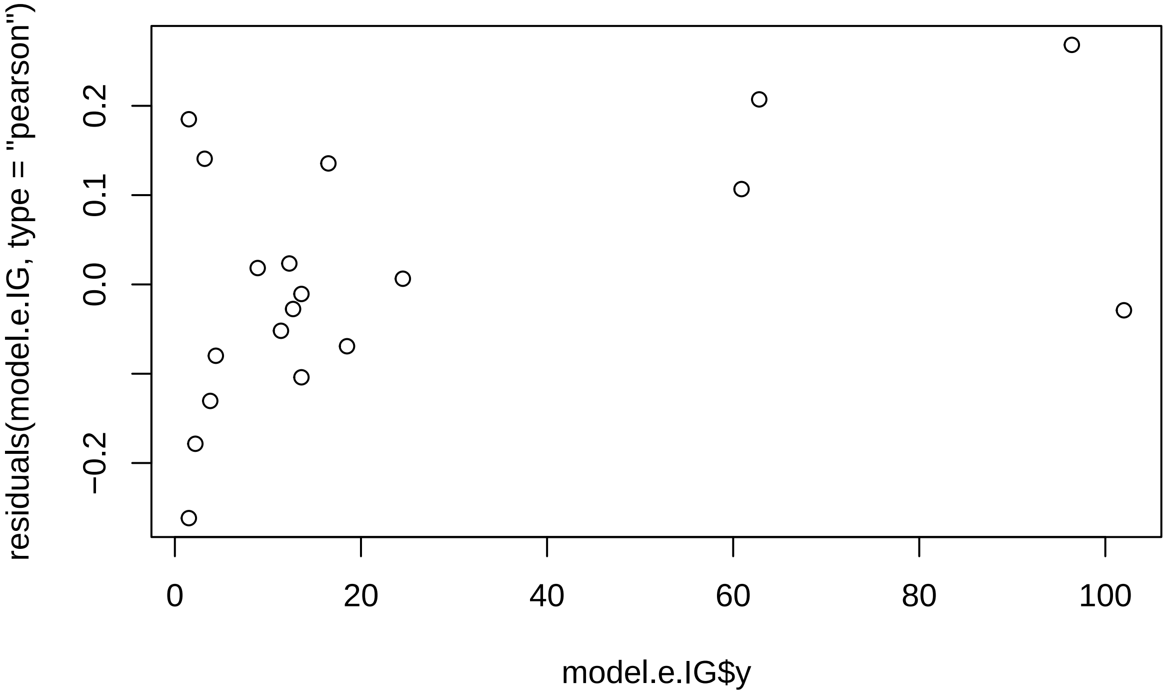
```
#plot(fitted(model.e.N, type="response"), residuals(model.e.N, type="pearson")^2)
#plot(fitted(model.e.G, type="response"), residuals(model.e.G, type="pearson")^2)
#plot(fitted(model.e.IG, type="response"), residuals(model.e.IG, type="pearson")^2)

plot(model.e.N$y, residuals(model.e.N, type="pearson"))
```



```
plot(model.e.G$y, residuals(model.e.G, type="pearson"))
```



```
plot(model.e.IG$y, residuals(model.e.IG, type="pearson"))
```



```
# Based on my analysis, I choose the inverse gaussian distrbution
#as it yield the highest p for normality test for Pearson's residual.
#IG distribtution also have the lowest MSE
```

```
#Problem 2
data2<-read.table("/Users/phamkhoa/Documents/university/3/stats_model_1/w2/ratstime.txt", sep="\t", dec
attach(data2)
```

```
#a
N.p2<-glm(time~factor(poison) + factor(treat), family = gaussian("identity"), data = data2)
G.p2<-glm(time~factor(poison) + factor(treat), family = Gamma("identity"), data = data2)
IG.p2<-glm(time~factor(poison) + factor(treat), family = inverse.gaussian("identity"), data = data2)

AIC(N.p2)
```

```
## [1] -33.21821
```

```
AIC(G.p2)
```

```
## [1] -62.23033
```

```
AIC(IG.p2)
```

```
## [1] -71.183
## MSE

mean(residuals(N.p2, type="response")^2)
```

```
## [1] 0.02189297
```

```
mean(residuals(G.p2, type="response")^2)
```

```
## [1] 0.02379278
```

```
mean(residuals(IG.p2, type="response")^2)
```

```
## [1] 0.02535005
```

```
shapiro.test(residuals(N.p2, type="pearson"))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(N.p2, type = "pearson")
## W = 0.92242, p-value = 0.003622
```

```
shapiro.test(residuals(G.p2, type="pearson"))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(G.p2, type = "pearson")
## W = 0.90006, p-value = 0.0006314
```

```
shapiro.test(residuals(IG.p2, type="pearson"))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(IG.p2, type = "pearson")
## W = 0.90504, p-value = 0.0009185
```

```
#I choose the model N.p2 (normal distribution identity link)
#as it have the smallest MSE and highest p for shapiro test
```

```
#b
p2b.identity<-glm(time~factor(poison) + factor(treat), family = Gamma("identity"), data = data2)
p2b.log<-glm(time~factor(poison) + factor(treat), family = Gamma("log"), data = data2)
p2b.inverse<-glm(time~factor(poison) + factor(treat), family = Gamma("inverse"), data = data2)

mean(residuals(p2b.identity, type="response")^2)
```

```
## [1] 0.02379278
```

```
mean(residuals(p2b.log, type="response")^2)
```

```
## [1] 0.01939997
```

```
mean(residuals(p2b.inverse, type="response")^2)
```

```
## [1] 0.01900666
```

```
# From the MSE value, choose inverse link with gramma distribution as it has the smallest MSE
```

```
#c
ig.H0<-glm(time~factor(poison) + factor(treat), family = inverse.gaussian("log"), data = data2)
ig.H1<-glm(time~factor(poison)*factor(treat), family = inverse.gaussian("log"), data = data2)
anova(ig.H0, ig.H1, test = "F")
```

```
## Analysis of Deviance Table
##
## Model 1: time ~ factor(poison) + factor(treat)
## Model 2: time ~ factor(poison) * factor(treat)
##   Resid. Df Resid. Dev Df Deviance      F Pr(>F)
## 1        42     4.6835
## 2        36     3.6418  6   1.0417 1.8462 0.1176
```

```
#Pr(>F) = 0.1176
```

```
#d
newdata<-data.frame(poison = "II", treat = "B")
pred<-predict(p2b.log, newdata=newdata, type="response")
pred
```

```
##         1
## 0.7304192
```

```
xf<-cbind(c(1,1,0,1,0,0))

Var.Yf<-summary(p2b.log)$dispersion
D.f<-pred
Var.ef<-Var.Yf+(D.f^2)*t(xf)%*%vcov(p2b.log)%*%xf

lower.yf<-pred-qnorm(0.9)*sqrt(Var.ef)
upper.yf<-pred+qnorm(0.9)*sqrt(Var.ef)

lower.yf
```

```
##           [,1]
## [1,] 0.4027708
```

```r
#[1,] lower 0.4027708
upper.yf
```

```
##          [,1]
## [1,] 1.058068
```

```r
#[1,] upper 1.058068
```

```r
#e
newdata<-data.frame(poison = c("I","II"), treat = c("D","B"))
pred<-predict(p2b.log, newdata=newdata, type="response")
pred
```

```
##         1         2
## 0.7028960 0.7304192
```

```r
x1<-cbind(c(1,0,0,0,0,1))
x2<-cbind(c(1,1,0,1,0,0))
Xf<-t(cbind(x1,x2))
k<-cbind(c(-1,1))

phi<-summary(p2b.log)$dispersion
Var.Y1f<-phi*(pred[1]^2)
Var.Y2f<-phi*(pred[2]^2)

D.f<-diag(pred)

Var.ef<-Var.Y1f+Var.Y2f+t(k)%*%D.f%*%Xf%*%vcov(p2b.log)%*%t(Xf)%*%D.f%*%k

lower.diff<-(pred[2]-pred[1])-qnorm(0.9)*sqrt(Var.ef)
upper.diff<-(pred[2]-pred[1])+qnorm(0.9)*sqrt(Var.ef)
lower.diff
```

```
##            [,1]
## [1,] -0.3167109
```

```r
upper.diff
```

```
##           [,1]
## [1,] 0.3717572
```

```r
#(-0.3167109, 0.3717572)
```